Original Paper

Biopsy-Free Prediction of Pathologic Type of Primary Nephrotic Syndrome Using a Machine Learning Algorithm

Cuifang Li^a Zhijiang Yao^b Minfeng Zhu^b Ben Lu^b Hui Xu^a

^aNephrology Department, Xiangya Hospital, Central South University, Changsha, Hunan, ^bHematology Department, The Third Xiangya Hospital, Central South University, Changsha, Hunan, China

Key Words

Primary nephrotic syndrome • Renal biopsy • Pathologic type • Machine learning algorithm

Abstract

Background/Aims: Renal biopsy is the gold standard to determine the pathologic type of primary nephrotic syndrome, which is critical for diagnosis, choice of treatment and evaluation of prognosis. However, in some cases, renal biopsy cannot be performed. **Methods:** To explore the possibility of predicting the histology type of primary nephrotic syndrome without the need for biopsy, we trained and validated a machine learning algorithm using data from 222 patients with biopsy-confirmed primary nephrotic syndrome treated at our hospital between May 2008 and January 2016. The model was then tested prospectively on another sample of 63 patients with biopsy-confirmed primary nephrotic syndrome. **Results:** Overall accuracy of prediction from the retrospective set of 222 patients was 62.2% across all types of nephrotic syndrome. The accuracy of model prediction for the prospectively collected dataset of 63 patients was 61.9%. The algorithm identified 17 of 33 variables as contributing strongly to type of renal pathology. **Conclusion:** To our knowledge, this is the first such application of machine learning to predict the pathologic type of primary nephrotic syndrome, which may be clinically useful by itself as well as helpful for guiding future efforts at machine learning-based prediction in other disease contexts.

© 2017 The Author(s)
Published by S. Karger AG, Basel

Introduction

Nephrotic syndrome, a hallmark of various types of glomerular diseases, involves proteinuria in excess of 3.5g per 24h, hypoalbuminemia (serum albumin <30 g/l), hyperlipidemia, and edema. It can occur as primary disease or secondary to systemic disorders such as systemic lupus erythematosus, allergic purpura, or amyloidosis. Primary nephrotic syndrome is usually categorized into five pathologic types, depending on morphological and structural changes of the glomeruli observed by light microscopy, electron microscopy

Hui Xu
and Ben Lu

Nephrology Department, Xiangya Hospital, Central South University, Changsha (China)
Tel. +86 15974199189, Tel. +86 18774959889
E-Mail xybenlu@csu.edu.cn; xuhuiye@csu.edu.cn

and immunofluorescence. These five types include minimal change glomerulopathy, IgA or non-IgA mesangial proliferative glomerulonephritis, membranoproliferative glomerulonephritis (including C3 glomerulonephritis and immune complex-mediated membranoproliferative glomerulonephritis), membranous nephropathy, and focal segmental glomerular sclerosis. Correctly determining the pathologic type of primary nephrotic syndrome is important for optimizing treatment and predicting prognosis [1].

The most rigorous way to type the pathology of nephrotic syndrome is through invasive kidney biopsy, which is also the case in several other primary and secondary kidney diseases [2]. However, some patients are unsuitable for renal biopsy, such as those with severe edema, a tendency to bleed or only one functioning kidney [3]. In addition, the invasive procedure is associated with life-threatening complications [4-6] and many hospitals in resource-limited settings lack the proper facilities to perform kidney biopsy. A reliable and non-invasive way to predict pathologic type of primary nephrotic syndrome based on easily obtained clinical data could provide a powerful alternative to biopsy.

Studies have already identified some correlations between patient characteristics and type of primary nephrotic syndrome. For example, age may be useful for differentiating patients at risk of minimal-change glomerulopathy (common in children) from those at risk of membranous nephropathy (rare in children) [7]. IgA-type mesangial proliferative glomerulonephritis is closely associated with gross hematuria, whereas minimal-change glomerulopathy features heavy proteinuria, generalized edema, and hypoalbuminemia [8-9]. Another distinguishing characteristic may be electrophoretic patterns of urinary proteins: minimal-change glomerulopathy is believed to involve dysfunction in the charge-selective barrier, while other types of primary nephrotic syndrome are often associated with dysfunction in the size-selective barrier. Given the complexity of glomerular diseases, it is unlikely that one or a few patient characteristics will be sufficient to reliably diagnose or predict the pathologic type of primary nephrotic syndrome. It is more likely that an array of characteristics need to be integrated, which suggests the need for a computational approach.

Machine learning algorithms are well suited to analyzing large, complex data sets. This approach does not start with a predefined model, rather it allows a model to emerge from existing data when the algorithm recognizes underlying patterns. Models are developed using a “training” dataset and then validated using a different dataset. This approach allows the diagnostic or predictive model to be continuously updated and refined as more data become available for training the algorithm. Already in use in several fields, machine learning is slowly entering the clinic. For example, Shouval et al. retrospectively analyzed clinical data using a machine learning algorithm to predict mortality at 100 days after allogeneic hematopoietic stem cell transplantation [10]. Their model predicted mortality significantly better than the European Group for Blood and Marrow Transplantation score.

In this work, we used machine learning to retrospectively analyze data from patients with primary nephrotic syndrome who underwent kidney biopsy, and we developed an algorithm to predict the pathologic type of primary nephrotic syndrome on the basis of routine clinical data obtained non-invasively. After training and validating the model, we tested it prospectively on a group of 63 patients with primary nephrotic syndrome. The results suggest the potential of using machine learning to predict the pathologic type of primary nephrotic syndrome for patients, especially those with only one functioning kidney, for whom renal biopsy is contra-indicated.

Materials and Methods

Patients

This study retrospectively analyzed medical records of 222 patients with primary nephrotic syndrome who were recruited between May 2008 and January 2016 from Xiangya Hospital of Central South University. The study also prospectively analyzed a separate group of 63 patients with primary nephrotic syndrome

treated at our hospital during the same period. All patients were diagnosed by professional renal physicians, and diagnosis was confirmed by renal biopsy. Patients were excluded from the study if they met any of the following criteria: (1) younger than 14 years; (2) diagnosed with nephrotic syndrome secondary to systemic lupus erythematosus, hepatitis B virus-associated nephritis, allergic purpura, diabetes, amyloid lesions, multiple myeloma, et al; (3) diagnosed with membranoproliferative glomerulonephritis; (4) renal biopsy results were unavailable. Patients screened for our study were diagnosed with membranoproliferative glomerulonephritis before publication of a revised classification scheme [11] according to which many cases of idiopathic membranoproliferative glomerulonephritis should be reclassified as complement component 3 glomerulopathy [12].

This study was approved by the Ethics Committee of Xiangya Hospital (approval no. 201303067). All methods were carried out in accordance with relevant guidelines and regulations, and informed consent was obtained from all subjects.

Data collection

Records were collected on the following patient characteristics: age, sex, medical history, blood pressure, results of routine blood analysis, urinary red cell morphology, serum albumin, creatinine, triglyceride, cholesterolin, blood glucose, fibrinogen, thyroid function, 24h urine protein excretion and results of urine protein electrophoresis. Renal biopsy results were used to assign each patient to a pathologic type of primary nephrotic syndrome.

Statistical analysis

Patient characteristics were reported as n (%) or as mean $\pm$ SD. A model predicting pathologic type of primary nephrotic syndrome from patient data was developed using the random forest algorithm [13] based on an ensemble of classification trees. In this method, each classification tree is built using a bootstrap sample of the data, with a random subset of variables chosen at each split to serve as the set of candidate variables. In this way, the random forest algorithm builds trees by combining bootstrap aggregation (bagging)[14] with random variable selection. Bagging can be particularly effective for obtaining a model from heterogeneous or complex data. The resulting trees show lower bias because they are allowed to “grow” fully (i.e. they are left “unpruned”), and the different trees do not correlate strongly with one another because of the combination of bagging and random variable selection. By averaging over a large ensemble of low-bias trees showing high variance but low inter-correlation, the random forest algorithm can generate results with minimal bias and variance.

Since quite different proportions of patients in our dataset showed certain physical or clinical characteristics, we wanted to reduce the risk that such imbalance would bias our model to favor variables that were better represented in our patients. Therefore we took sample size

Table 1. Characteristics of patients with primary nephrotic syndrome used to train and validate the predictive model. Values are reported as n or as mean $\pm$ SD

Characteristic	Value
n	222
Males:females	135:87
Age, yr (range)	40.04 $\pm$ 15.90 (14-72)
24h urine volume, L	1.69 $\pm$ 0.84
24h urinary protein, g	4.90 $\pm$ 4.08
Serum albumin, g/l	21.72 $\pm$ 6.75
serum globulin, g/l	21.53 $\pm$ 3.48
Creatinine, mol/l	110.66 $\pm$ 88.84
Uric acid, mol/l	395.53 $\pm$ 212.22
Blood urea nitrogen, mmol/l	7.53 $\pm$ 5.27
Triglyceride, mmol/l	2.51 $\pm$ 1.48
Cholesterolin, mmol/l	8.58 $\pm$ 3.01
Free triiodothyronine, pmol/l	2.94 $\pm$ 1.25
Free thyroxine, pmol/l	10.37 $\pm$ 6.01
Thyroid-stimulating hormone, mIU/L	4.52 $\pm$ 8.47
Platelet count, 10 ⁹ /l	236.31 $\pm$ 71.87
Hemoglobin, g/l	129.52 $\pm$ 20.55
Fibrinogen, g/l	7.95 $\pm$ 34.19
Complement C3, mg/l	1038.54 $\pm$ 283.51
Immunoglobulin G, mg/l	4953.47 $\pm$ 2410.06
Immunoglobulin A, mg/l	1915.36 $\pm$ 999.92
Immunoglobulin M, mg/l	1270.15 $\pm$ 702.27
Erythrocyte sedimentation rate, mm/h	39.85 $\pm$ 26.06
C-reactive protein, mg/l	5.98 $\pm$ 16.09
Blood glucose, mmol/l	4.78 $\pm$ 1.54
Systolic pressure, mmHg	130.25 $\pm$ 18.40
Diastolic pressure, mmHg	81.90 $\pm$ 12.10
Immune protein electrophoresis, %	
Macromolecules	11.08 $\pm$ 14.33
Mid-sized molecules	87.41 $\pm$ 16.24
Small molecules	1.50 $\pm$ 7.46
Type of renal pathology, n	
Minimal-change glomerulopathy	23
Focal segmental glomerulosclerosis	19
Membranous nephropathy	88
Non-IgA mesangial proliferative glomerulonephritis	64
IgA nephropathy	28

into account when bootstrapping the sample, ensuring similar numbers of patients positive or negative in each bootstrap sampling. Sample size was set to different values depending on the particular dataset.

Results

Patient characteristics

Baseline characteristics of the 222 patients whose data were used to train and validate the predictive model are shown in Table 1. The ratio of males (135) to females (87) was 1.55:1. Age ranged from 14 to 72 years, and average age was 40.04 ± 15.90 years. All pathologic types of primary nephrotic syndrome were represented in our sample, except for membranoproliferative glomerulonephritis (Table 1); the largest proportion had membranous nephropathy (39.64%). This is consistent with previous studies of patients from central and northern China [15-16]. The smallest proportion of patients had focal segmental glomerulosclerosis (8.56%). Patients with minimal-change nephropathy were, on average, younger than those with membranous nephropathy, which is consistent with previous reports [17].

Model development

Using the training set of retrospective data on 222 patients, we used the random forest algorithm to develop a predictive model (Fig. 1). The random forest algorithm identified 17 of 33 variables as contributing strongly to pathologic type of primary nephrotic syndrome (Fig. 2).

Model validation

Based on the training dataset of 222 patients, the model predicted the pathologic type of primary nephrotic syndrome with an accuracy of 62.2% across all types. The accuracy for most types was roughly similar, but quite poor for focal segmental glomerulosclerosis: membranous nephropathy, 76.1%; non-IgA mesangial proliferative glomerulonephritis, 64.1%; IgA nephropathy, 57.1%; minimal-change disease, 52.2%; and focal segmental glomerulosclerosis, 10.5% (Table 2). Based on the validation dataset from a separate group of 63 patients analyzed prospectively, overall accuracy for predicting pathologic type was 61.9% (Table 3).

Table 2. Comparison of biopsy-determined pathologic types of primary nephrotic syndrome (columns). and model-predicted types (rows) using the retrospective dataset of 222 patients. MCD--minimal-change glomerulopathy, FSGS--focal segmental glomerulosclerosis, MN--membranous nephropathy, Non-IgAN--non-immunoglobulin A nephropathy, IgAN--immunoglobulin A nephropathy

	MCD (n=23)	FSGS (n=19)	MN (n=88)	Non-IgAN (n=64)	IgAN (n=28)
MCD	12	3	4	3	0
FSGS	1	2	1	0	1
MN	4	7	67	17	3
Non-IgAN	4	3	14	41	8
IgAN	2	4	2	3	16

Table 3. Comparison of biopsy-determined pathologic types of primary nephrotic syndrome and model-predicted types using the prospective dataset of 63 patients. MCD--minimal-change glomerulopathy, FSGS--focal segmental glomerulosclerosis, MN--membranous nephropathy, Non-IgAN--non-immunoglobulin A nephropathy, IgAN--immunoglobulin A nephropathy, CV--coefficient of variation

	Overall	MCD	FSGS	MN	Non-IgAN	IgAN
CV	0.622	0.522	0.105	0.761	0.641	0.571
Predict	61.9	85.7	0	63.6	60	80

Fig. 1. Flow chart of the random forest method to predict pathologic type of primary nephrotic syndrome from patient and clinical data. The random forest classification model comprises 1000 trees. At each tree node or branch, the algorithm maximizes the separation of all instances, until the tree is able to classify all instances. At this point, the node becomes a terminal node. On the top left side of the Fig.(from DT1 to DT1000) represents the final random forest modeling graphics, including a total of 1000 trees. On the right side of the Fig. (Decision Tree) represents the magnified diagram of a tree in the 1000 trees. Class 1 to class 5 respectively represent minimal-change glomerulopathy; focal segmental glomerulosclerosis; membranous nephropathy; non-immunoglobulin A nephropathy and immunoglobulin A nephropathy.

Fig. 2. Relative importance of patient and clinical characteristics identified by the random forest algorithm for generating a model to predict pathologic type of primary nephrotic syndrome. Of the 33 variables initially included in the model, only the 17 most important are shown.

Discussion

Here we demonstrate the feasibility of a machine learning-based approach to predict features of primary nephrotic syndrome pathology in patients for whom invasive biopsy is contra-indicated.

Our work vindicates a previous suggestion that certain salient features of the patient, such as demographic, clinical or laboratory characteristics, may allow the clinician to predict pathologic type without the need for biopsy [18]. To our knowledge, this is the first application of machine learning to predict the pathologic type of primary nephrotic syndrome. The clinical data required for prediction are easily obtained through routine clinical procedures. The fact that we have been able to achieve reasonably good predictions based on a complex clinical dataset suggests the feasibility of using machine learning to predict diagnosis and perhaps prognosis in other disease contexts as well, including secondary nephrotic syndromes.

In our group of patients, the disease disproportionately affected males (1.55:1). This is consistent with previous reports of male gender bias of primary nephrotic syndrome. Indeed, our predictive model identified gender as having a substantial effect on the pathologic type.

In addition, with the economic level and people's health consciousness improved, the average age of patients with renal biopsy have increased. In this study, more than 50 years old were over 30 percent of the total number of patients with renal biopsy, which is consistent with previous studies of patients from central and northern China [19].

Environmental and genetic differences likely help explain the different prevalences observed among the four pathologic types of primary nephrotic syndrome in our patients [20-21]. Membranous nephropathy was the most common type in our patients, as reported in other Chinese populations [22-23], while focal segmental glomerulosclerosis was the least common type.

While our results suggest the potential of our machine learning approach, they should be interpreted with caution because of several limitations. The fact that we trained and validated the model on a retrospective dataset raises the risk of measurement bias [24]. Nevertheless, we showed good model performance with a prospectively collected dataset independent of the validation dataset, and our results are consistent with previous reports of correlations between clinical parameters and renal pathology [25]. At the same time, the quite poor accuracy of predicting focal segmental glomerulosclerosis indicates that the model shows room for improvement, or that it may be limited only to predicting certain types of primary nephrotic syndrome well. The poor accuracy with focal segmental glomerulosclerosis may reflect the small number of patients with that type in our dataset. It may be possible to improve this accuracy by enlarging the training dataset. It is also possible that predicting this type of nephrotic syndrome may be intrinsically difficult because of its heterogeneity. The Columbia Classification distinguishes five variants of this pathologic type: collapsing, tip, cellular, perihilar, and "not otherwise specified" [26]. In addition, the classification of focal segmental glomerulosclerosis depends on the percentage of glomeruli affected, the number of serial sections, and the size of segmental lesions [27]. Therefore it may be difficult to predict focal disease on the basis of only a few clinical features.

In conclusion, we have described a machine learning model based on the random forest algorithm to predict the pathologic type of primary nephrotic syndrome based on readily available, non-invasive patient and clinical data. The random forest approach may prove fruitful for predictive diagnosis of other diseases as well, including secondary nephrotic syndromes such as diabetic nephropathy and systemic lupus erythematosus.

Disclosure Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

- 1 MacHardy N, Miles PV, Massengill SF, Smoyer WE, Mahan JD, Greenbaum L, Massie S, Yao L, Nagaraj S, Lin JJ, Wigfall D, Trachtman H, Hu Y, Gipson DS: Management of childhood onset nephrotic syndrome. *Pediatrics* 2009;124:747-757.
- 2 Visconti L, Cernaro V, Ricciardi CA, Lacava V, Pellicanò V, Lacquaniti A, Buemi M, Santoro D: Renal biopsy: Still a landmark for the nephrologist. *World J Nephrol* 2016;5:321.
- 3 Madaio MP: Renal biopsy. *Kidney Int* 1990;38:529-543.
- 4 Huang CC, Kuo CC, Chen YM: The incidence of fatal kidney biopsy. *Clin Nephrol* 2011;76:256-258.
- 5 Korbet SM, Volpini KC, Whittier WL: Percutaneous Renal Biopsy of Native Kidneys: A Single-Center Experience of 1,055 Biopsies. *Am J Nephrol* 2014;39:153.
- 6 Feneberg R, Schaefer F, Zieger B, Waldherr R, Mehls O, Schärer K: Percutaneous Renal Biopsy in Children: A 27-Year Experience. *Nephron* 1998;79:438-446.
- 7 Eddy AA, Symons JM: Nephrotic syndrome in childhood. *Lancet* 2003;362:629.
- 8 Kim JK, Kim JH, Lee SC, Kang EW, Chang TI, Moon SJ, Yoon SY, Yoo TH, Kang SW, Choi KH, Han DS, Kie JH, Lim BJ, Jeong HJ, Han SH: Clinical features and outcomes of IgA nephropathy with nephrotic syndrome. *Clin J Am Soc Nephrol* 2012;7:427-436.
- 9 Mao HX, Yi ZW, Liang L, He YW, Dang XQ, Wu XC, He XJ, Mo SH: Analysis of 94 cases of IgA nephropathy in children. *Zhong Nan Da Xue Xue Bao Yi Xue Ban* 2007;32:174-178.
- 10 Shouval R, Labopin M, Bondi O, Mishan-Shamay H, Shimoni A, Ciceri F, Esteve J, Giebel S, Gorin NC, Schmid C, Polge E, Aljurf M, Kroger N, Craddock C, Bacigalupo A, Cornelissen JJ, Baron F, Unger R, Nagler A, Mohty M: Prediction of Allogeneic Hematopoietic Stem-Cell Transplantation Mortality 100 Days After Transplantation Using a Machine Learning Algorithm: A European Group for Blood and Marrow Transplantation Acute Leukemia Working Party Retrospective Data Mining Study. *J Clin Oncol* 2015;33:3144-3151.
- 11 Sethi S, Haas M, Markowitz GS, D'Agati VD, Rennke HG, Jennette JC, Bajema IM, Alpers CE, Chang A, Cornell LD, Cosio FG, Fogo AB, Glassock RJ, Hariharan S, Kambham N, Lager DJ, Leung N, Mengel M, Nath KA, Roberts IS, et al: Mayo Clinic/Renal Pathology Society Consensus Report on Pathologic Classification, Diagnosis, and Reporting of GN. *J Am Soc Nephrol* 2016;27:1278-1287.
- 12 Kawasaki Y, Kanno S, Ono A, Suzuki Y, Ohara S, Sato M, Suyama K, Hashimoto K, Hosoya M: Differences in clinical findings, pathology, and outcomes between C3 glomerulonephritis and membranoproliferative glomerulonephritis. *Pediatr Nephrol* 2016;31:1091-1099.
- 13 Breiman L: Bagging predictors. *Machine learning* 1996;24:123-140.
- 14 Breiman L: Random forests. *Machine learning* 2001;45:5-32.
- 15 Chu F, Chen G, Liu Y: Pathological patterns of primary nephrotic syndrome in Central China: a retrospective study of 627 cases. *Ren Fail* 2014;36:514-519.
- 16 Zhou FD, Shen HY, Chen M, Liu G, Zou WZ, Zhao MH, Wang HY: The renal histopathological spectrum of patients with nephrotic syndrome: an analysis of 1523 patients in a single Chinese centre. *Nephrol Dial Transplant* 2011;26:3419-3420.
- 17 Waldman M, Crew RJ, Valeri A, Busch J, Stokes B, Markowitz G, D'Agati V, Appel G: Adult minimal-change disease: clinical characteristics, treatment, and outcomes. *Clin J Am Soc Nephrol* 2007;2:445-453.
- 18 Grace M, Thomas V: Clinicopathological correlation of primary nephrotic syndrome in adults. *Saudi J Kidney Dis Transpl* 2012;23:1292-1293.
- 19 Chen CM, Xia YC: The relationship between pathologic type and hormone efficacy in adult primary nephrotic syndrome. *Guangdong Medical Journal* 2015;36:1040-1042.

- 20 Han DU, Feng HU, Li DU: A summary on the pathological data from 480 cases of renal biopsy. *Pract J Med Pharm* 2010; 27:390-392.
- 21 Yu S: Analysis on Clinical Pathological Data of 5000 Cases Undergoing Renal Biopsy in a Hospital in Xi'an. *Chinese J Prevent Contr Chronic Dis* 2011;19:289-291.
- 22 Haas M, Meehan SM, Karrison TG, Spargo BH: Changing etiologies of unexplained adult nephrotic syndrome: A comparison of renal biopsy findings from 1976–1979 and 1995–1997. *Am J Kidney Dis* 1997;30:621-631.
- 23 Dragovic D, Rosenstock JL, Wahl SJ, Panagopoulos G, DeVita MV, Michelis MF: Increasing incidence of focal segmental glomerulosclerosis and an examination of demographic patterns. *Clin Nephrol* 2005;63:1-7.
- 24 Ward RA, Brier ME: Retrospective analyses of large medical databases: what do they tell us?. *J Am Soc Nephrol* 1999;10:429-432.
- 25 Lauer MS, Sr DR: The randomized registry trial-the next disruptive technology in clinical research? *N Engl J Med* 2013;369:1579-1581.
- 26 Stokes MB, D'Agati VD: Morphologic variants of focal segmental glomerulosclerosis and their significance. *Adv Chronic Kidney Dis* 2014;21:400-407.
- 27 Han MH, Kim YJ: Practical Application of Columbia Classification for Focal Segmental Glomerulosclerosis. *Biomed Res Int* 2016;2016:9375753.